Estimation of environment-related properties of chemicals for design of sustainable processes: Development of group-contribution+ (GC+) models and uncertainty analysis

Hukkerikar, Amol; Kalakul, Sawitree; Sarup, Bent; Young, Douglas M.; Sin, Gürkan; Gani, Rafiqul

Published in:
Journal of Chemical Information and Modeling

Link to article, DOI:
10.1021/ci300350r

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Estimation of Environment-Related Properties of Chemicals for Design of Sustainable Processes: Development of Group-Contribution$^+$ (GC$^+$) Property Models and Uncertainty Analysis

Amol Shivajirao Hukkerikar,[†] Sawitree Kalakul,[‡] Bent Sarup,[§] Douglas M. Young,[⊥] Gürkan Sin,[†] and Rafiqul Gani*,[†]

[†]Computer Aided Process-Product Engineering Center (CAPEC), Department of Chemical and Biochemical Engineering, Technical University of Denmark, DK-2800, Kgs. Lyngby, Denmark

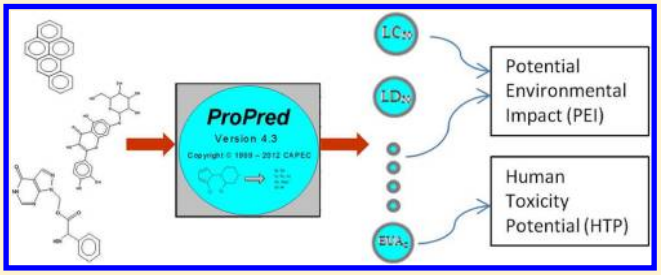
[‡]The Petroleum and Petrochemical College, Chulalongkorn University, Bangkok 10330, Thailand

[§]Vegetable Oil Technology Business Unit, Alfa Laval Copenhagen A/S, Maskinvej 5, DK-2860, Søborg, Denmark

[⊥]U.S. EPA, National Risk Management Research Laboratory, Cincinnati, Ohio 45268, United States

Supporting Information

ABSTRACT: The aim of this work is to develop group-contribution$^+$ (GC$^+$) method (combined group-contribution (GC) method and atom connectivity index (CI) method) based property models to provide reliable estimations of environment-related properties of organic chemicals together with uncertainties of estimated property values. For this purpose, a systematic methodology for property modeling and uncertainty analysis is used. The methodology includes a parameter estimation step to determine parameters of property models and an uncertainty analysis step to establish statistical information about the quality of parameter estimation, such as the parameter covariance, the standard errors in predicted properties, and the confidence intervals. For parameter estimation, large data sets of experimentally measured property values of a wide range of chemicals (hydrocarbons, oxygenated chemicals, nitrogenated chemicals, poly functional chemicals, etc.) taken from the database of the US Environmental Protection Agency (EPA) and from the database of USEtox is used. For property modeling and uncertainty analysis, the Marrero and Gani GC method and atom connectivity index method have been considered. In total, 22 environment-related properties, which include the fathead minnow 96-h LC_{50}, *Daphnia magna* 48-h LC_{50}, oral rat LD_{50}, aqueous solubility, bioconcentration factor, permissible exposure limit (OSHA-TWA), photochemical oxidation potential, global warming potential, ozone depletion potential, acidification potential, emission to urban air (carcinogenic and noncarcinogenic), emission to continental rural air (carcinogenic and noncarcinogenic), emission to continental fresh water (carcinogenic and noncarcinogenic), emission to continental seawater (carcinogenic and noncarcinogenic), emission to continental natural soil (carcinogenic and noncarcinogenic), and emission to continental agricultural soil (carcinogenic and noncarcinogenic) have been modeled and analyzed. The application of the developed property models for the estimation of environment-related properties and uncertainties of the estimated property values is highlighted through an illustrative example. The developed property models provide reliable estimates of environment-related properties needed to perform process synthesis, design, and analysis of sustainable chemical processes and allow one to evaluate the effect of uncertainties of estimated property values on the calculated performance of processes giving useful insights into quality and reliability of the design of sustainable processes.



INTRODUCTION

Currently, there is a great deal of interest in the development of computer aided methods and tools for the process synthesis, design, and analysis of sustainable processes. The design of sustainable processes requires the satisfying of various conditions (or constraints) such as, increased productivity, minimum energy consumption, reduction in raw materials, recovery of products, and minimum generation of pollution.[1] This task can be effectively accomplished by using a chemical process simulator (to perform mass and energy balances for the concerned process) together with the waste reduction (WAR) algorithm[2,3] to obtain a quantitative measure of the potential environmental impact (PEI) which, as part of the life cycle assessment (LCA) of process synthesis and design, contributes to identifying sustainable processing paths and design alternatives. The PEI is a relative measure of the potential for a chemical to have an adverse effect on human health and the environment. Several

Received: July 27, 2012
Published: October 7, 2012

studies in the literature[4−6] have reported the application of the WAR algorithm for generating sustainable process design alternatives and deciding on sustainable process designs that are environmentally friendly and economically attractive. In the WAR algorithm, the total PEI of a process is evaluated based on the following eight categories of potential impacts: (i) human toxicity potential by ingestion, calculated using oral rat LD_{50}; (ii) human toxicity potential by exposure both dermal and inhalation, calculated using permissible exposure limits (OSHA-TWA); (iii) terrestrial toxicity potential, calculated using oral rat LD_{50}; (iv) aquatic toxicity potential, calculated using fathead minnow 96-h LC_{50}; (v) global warming potential; (vi) ozone depletion potential; (vii) photochemical oxidation potential; and (viii) acidification potential. Therefore, the basis for the quantification of PEI is a set of environment-related properties (such as fathead minnow 96-h LC_{50}, oral rat LD_{50}, global warming potential, etc.) of chemical substances involved in the process. The USEtox model[7] is an environment model for characterization of human and ecotoxicological impacts in life cycle impact assessment (LCIA), and comparative risk assessment (CRA) and is designed to describe the fate, exposure, and effects of chemicals.[7,8] The USEtox model calculates characterization factors for carcinogenic impacts, noncarcinogenic impacts, and total impacts (carcinogenic + noncarcinogenic) based on the chemical emissions to urban air, rural air, freshwater, seawater, agricultural soil, and/or natural soil. The definition of each environment-related property considered in this work is given in Table 1. For many chemicals of interest, the experimental data of environment-related properties is not available since the measurement of these properties is extremely time-consuming and expensive. Also, processes that deal with the synthesis of new chemicals require a suitable property prediction method in order to obtain reliable estimates of environment-related properties of new chemicals. A review article by Boethling et al.[9] discusses available experimental data sources and various estimation methods including group-contribution (GC) methods, methods based on quantitative structure−property relationships (QSPR), and correlation equations, to name a few for obtaining values of environment-related properties of chemicals.

For the estimation of properties of organic chemicals, GC methods such as those reported by Joback and Reid,[13] Lydersen,[14] Constantinou and Gani,[15] and Marrero and Gani[16] have been widely employed to obtain the needed property values since these methods provide the advantage of quick estimates without requiring substantial computational work. In GC methods, the property of a chemical is a function of structurally dependent parameters, which are determined as a function of the frequency of the groups representing the chemical and their contributions. Among GC methods for estimation of properties of chemicals, the Marrero and Gani (MG) method[16] is well-known. The MG method allows estimation of properties based exclusively on the molecular structure of the chemical and exhibits a good accuracy and a wide range of applicability covering chemical, biochemical, and environment-related chemicals. Note that for reliable estimation of properties of chemicals using a GC method, the user needs (i) a property model; (ii) group definitions (model parameters of the selected property model) and their contributions; and (iii) a tool to quantify uncertainties (prediction errors) of estimated property values in order to check the quality (reliability) of estimation. In many cases, a user may come across a situation where the selected GC model may not have all the model parameters (that is, groups describing the structure of a given organic chemical) and/or parameter values (that is, group contributions) needed for the estimation of property of that chemical. This issue is due to (i) lack of necessary experimental data of properties of organic chemicals in the parameter estimation step and (ii) lack of necessary group definitions required to describe the complete structure of wide range of organic chemicals. In such situations where the molecular structure of a given organic chemical is not completely described by any of the available groups, the atom connectivity index (CI) method can be employed to create the missing groups and/or to predict missing group contributions. These created missing groups and/or predicted missing group contributions obtained using the CI method are then used in the GC method together with other available groups of the GC method and their group contributions in order to estimate the property of that organic chemical. This combined approach is known as the group-contribution$^+$ (GC^+) method.[17] For example, let us consider the organic chemical methane, dimethoxy-. It can be represented by following Marrero and Gani GC method groups: CH_3 and CH_3O. Now let us assume that the user needs to estimate oral rat LD_{50} for this chemical, and let us further assume that contribution of the CH_3 group is available but the contribution of CH_3O group is not available in the list of model parameter values for oral rat LD_{50}. In this case, the user cannot use the Marrero and Gani GC method to estimate oral rat LD_{50} of methane, dimethoxy-. To overcome this limitation, the CI method is used to predict the missing contribution of the CH_3O group. This predicted contribution of CH_3O together with already available contribution of CH_3 group is then used in the Marrero and Gani GC method to estimate oral rat LD_{50} of methane, dimethoxy-. Note that for illustration purpose, the Marrero and Gani GC method is considered here. However, the CI method can be combined with any available GC method for the purpose of creating missing groups and/or predicting missing group contributions.

There are numerous LCA software tools available (for example, SimaPro,[18] GaBi,[19] etc.) for quantification of potential impact that the processes would have on the environment on average. Most of these tools have built-in databases containing properties of chemicals needed for the environmental-impact analysis. However, for chemicals that are not included in the database, a suitable property prediction method is necessary to obtain the needed environment-related property values which will allow one to perform synthesis, design, and analysis of sustainable chemical processes. For the estimation of fathead minnow 96-h LC_{50} and aqueous solubility, various GC-methods have been developed. Martin and Young[20] developed a GC method to correlate the acute toxicity (96-h LC_{50}) to the fathead minnow using 397 organic chemicals based on the multilinear regression and computational neural networks approach for the parameter estimation. Casalegno et al.[21] used a diatomic fragment approach based GC method to correlate the acute toxicity (96-h LC_{50}) of 607 organic chemicals. For the estimation of aqueous solubility, Marrero and Gani[22] developed a GC method using a three-level parameter estimation approach (with a data set of 2087 organic chemicals used for the regression purpose), and this method requires only molecular structural information for the estimation of aqueous solubility. There are several other GC methods available for estimation of aqueous solubility (those of Klopman and Zhu[23] and Kühne et al.[24]). For the estimation of oral rat LD_{50} and bioconcentration factor (BCF), the more common approach has been to employ correlation equations (for example, bioconcentration factor for a chemical is estimated using known value of its octanol/water

Table 1. Definitions of Environment-Related Properties Considered in this Work

sl. no.	property	definition
1	fathead minnow 96-h LC_{50} (LC_{50}(FM)) in moles per liter	the fathead minnow LC_{50} end point represents the concentration in water which kills half of fathead minnow (*Pimephales promelas*) in 4 days (96 h)[10]
2	*Daphnia magna* 48-h LC_{50} (LC_{50}(DM)) in moles per liter	the *Daphnia magna* LC_{50} end point represents the concentration in water which kills half of *Daphnia magna* (a water flea) in 48 h[10]
3	oral rat LD_{50} (LD_{50}) in moles per kilogram	the oral rat LD_{50} end point represents the amount of the chemical (mass of the chemical per body weight of the rat) which when orally ingested kills half of rats[10]
4	aqueous solubility (Log W_S) in milligrams per liter	aqueous solubility is defined as the amount of a chemical that will dissolve in liquid water to form a homogeneous solution[10]
5	bioconcentration factor (BCF)	the bioconcentration factor is defined as the ratio of the chemical concentration in biota as a result of absorption via the respiratory surface to that in water at steady state[10]
6	permissible exposure limit (OSHA-TWA) in moles per m^3	the permissible exposure limit (OSHA-TWA) is a legal limit in the United States for exposure of an employee to a chemical substance or physical agent; it is usually given as a time-weighted average (TWA); a TWA is the average exposure over a specified period of time, usually a nominal 8 h; this means that, for limited periods, a worker may be exposed to concentrations higher than the permissible exposure limit, so long as the average concentration over 8 h remains lower[11]
7	photochemical oxidation potential (PCO)	the photochemical oxidation potential is the result of reactions that take place between nitrogen oxides and volatile organic components exposed to UV radiation. It is expressed using a reference substance such as ethylene[12]
8	global warming potential (GWP)	the global warming potential is calculated as a sum of emissions of the greenhouse gases (CO_2, N_2O, CH_4, and VOCs) multiplied by their respective GWP factors[12]
9	ozone depletion potential (ODP)	the ozone depletion potential is defined as the ozone depletion produced by a unit of the gas converted into ozone depletion values produced by the reference substance trichlorofluoromethane[12]
10	acidification potential (AP)	the acidification potential is a measure of the disposition of a unit of the mass of a component to release H^+ protons, expressed in terms of the H^+ potential of the reference substance SO_2[12]
11	emission to urban air (EUA_C) in cases per kilogram emitted (carcinogenic)	estimated increase in morbidity (carcinogenic) in the total human population per unit mass of a chemical emitted in the urban air compartment (that is, emission to higher population density)[7]
12	emission to urban air (EUA_{NC}) in cases per kilogram emitted (noncarcinogenic)	estimated increase in morbidity (noncarcinogenic) in the total human population per unit mass of a chemical emitted in the urban air compartment (that is, emission to higher population density)[7]
13	emission to continental rural air (ERA_C) in cases per kilogram emitted (carcinogenic)	estimated increase in morbidity (carcinogenic) in the total human population per unit mass of a chemical emitted in the rural air compartment (that is, emission to lower population density, lower stratosphere, and upper troposphere)[7]
14	emission to continental rural air (ERA_{NC}) in cases per kilogram emitted (noncarcinogenic)	estimated increase in morbidity (noncarcinogenic) in the total human population per unit mass of a chemical emitted in the rural air compartment (that is, emission to lower population density, lower stratosphere, and upper troposphere)[7]
15	emission to continental fresh water (EFW_C) in cases per kilogram emitted (carcinogenic)	estimated increase in morbidity (carcinogenic) in the total human population per unit mass of a chemical emitted in the fresh water compartment (that is, emission to lakes, rivers, and groundwater)[7]
16	emission to continental fresh water (EFW_{NC}) in cases per kilogram emitted (noncarcinogenic)	estimated increase in morbidity (noncarcinogenic) in the total human population per unit mass of a chemical emitted in the fresh water compartment (that is, emission to lakes, rivers, and groundwater)[7]
17	emission to continental seawater (ESW_C) in cases per kilogram emitted (carcinogenic)	estimated increase in morbidity (carcinogenic) in the total human population per unit mass of a chemical emitted in the seawater compartment[7]
18	emission to continental seawater (ESW_{NC}) in cases per kilogram emitted (noncarcinogenic)	estimated increase in morbidity (noncarcinogenic) in the total human population per unit mass of a chemical emitted in the seawater compartment[7]
19	emission to continental natural soil (ENS_C) in cases per kilogram emitted (carcinogenic)	estimated increase in morbidity (carcinogenic) in the total human population per unit mass of a chemical emitted in the natural soil compartment (that is, emission to forestry and industrial soil)[7]
20	emission to continental natural soil (ENS_{NC}) in cases per kilogram emitted (noncarcinogenic)	estimated increase in morbidity (noncarcinogenic) in the total human population per unit mass of a chemical emitted in the natural soil compartment (that is, emission to forestry and industrial soil)[7]
21	emission to continental agri-cultural soil (EAS_C) in cases per kilogram emitted (carcinogenic)	estimated increase in morbidity (carcinogenic) in the total human population per unit mass of a chemical emitted in the agricultural soil compartment[7]

dx.doi.org/10.1021/ci300350r | J. Chem. Inf. Model. 2012, 52, 2823−2839

Table 1. continued

sl. no.	property	definition
22	emission to continental agricultural soil (EAS_{NC}) in cases per kilogram emitted (noncarcinogenic)	estimated increase in morbidity (noncarcinogenic) in the total human population per unit mass of a chemical emitted in the agricultural soil compartment[7]

partition coefficient). Moreover, Martin et al.[25] have developed a hierarchical clustering technique to predict a variety of end points, including oral rat LD_{50}, BCF, aqueous solubility, and fathead minnow LC_{50} that combines group contributions with descriptors from graph theory. Software platforms have been developed both in the U.S. (US EPA 2012[10,26]) and in Europe (Isitutio Mario Negri 2012[27]) to predict these same end points. The application range and capability of these estimation equations is limited by the availability of the required property values. To the best of our knowledge, there are no GC methods reported in the literature for the estimation of the following environment-related properties: permissible exposure limit (OSHA-TWA), global warming potential, photochemical oxidation potential, ozone depletion potential, acidification potential, emission to urban air (carcinogenic and non-carcinogenic), emission to continental rural air (carcinogenic and noncarcinogenic), emission to continental fresh water (carcinogenic and noncarcinogenic), emission to continental seawater (carcinogenic and noncarcinogenic), emission to continental natural soil (carcinogenic and noncarcinogenic), and emission to continental agricultural soil (carcinogenic and noncarcinogenic). In addition to the accurate estimation of environment-related properties, it is also important to know the uncertainties (for example, prediction error in terms of 95% confidence interval) of the estimated property values that arise due to uncertainties of the regressed model parameters as well as due to approximate nature of the selected property model. With this information, it is possible to evaluate the effect of these uncertainties on the calculated potential impact that the processes would have on the environment and to verify the quality and reliability of the model-based design of sustainable processes.

Motivated by the preceding literature review and by the need of reliable estimation of environment-related properties in synthesis, design, and analysis of sustainable processes, this work aims to develop property prediction models based on the GC^+ approach (combined GC method and CI method) to provide reliable estimates of environment-related properties together with uncertainties of the estimated property values. For this purpose, a systematic methodology for property modeling and uncertainty analysis developed by Hukkerikar et al.[28] is used. The methodology includes a parameter estimation step to determine parameters (group/atom contributions, adjustable parameters, and a universal parameter) of property models and an uncertainty analysis step to establish statistical information about the quality of parameter estimation, such as the parameter covariance, the standard errors in predicted properties, and the confidence intervals. For property modeling with a GC method, the MG method[16] has been considered. For property modeling with a CI method, the models proposed by Gani et al.[17] have been considered. For parameter estimation, large data sets of experimentally measured property values of wide range of chemicals taken from the database of US Environmental Protection Agency (EPA)[10] and from the database of USEtox[7] is used. In total 22 environment-related properties, which include the fathead minnow 96-h LC_{50} (LC_{50}(FM)), *Daphnia magna* 48-h LC50 (LC_{50}(DM)), oral rat LD_{50}, aqueous solubility (Log W_s), bioconcentration factor (BCF), permissible exposure limit (PEL(OSHA-TWA)), photochemical oxidation potential (PCO), global warming potential (GWP), ozone depletion potential (ODP), acidification potential (AP), emission to urban air (carcinogenic (EUA_C) and noncarcinogenic (EUA_{NC})), emission to continental rural air (carcinogenic (ERA_C) and

noncarcinogenic (ERA_{NC})), emission to continental fresh water (carcinogenic (EFW_C) and noncarcinogenic (EFW_{NC})), emission to continental seawater (carcinogenic (ESW_C) and noncarcinogenic (ESW_{NC})), emission to continental natural soil (carcinogenic (ENS_C) and noncarcinogenic (ENS_{NC})), emission to continental agricultural soil (carcinogenic (EAS_C), and noncarcinogenic (EAS_{NC})) have been modeled and analyzed.

The paper first gives a brief overview of the methodology for property modeling and uncertainty analysis; followed by model performance statistics; and finally, application of the developed property models for estimation of environment-related properties. Tables containing list of property model parameters together with parameter values, due to their large size, are provided as Supporting Information.

METHODS AND TOOLS FOR PROPERTY MODELING AND UNCERTAINTY ANALYSIS

MG Group-Contribution Method. In the MG method,[16] the property estimation is performed at three levels. The first level has a large set of simple groups that allow for the representation of a wide variety of organic chemicals. However, these groups only partially capture the proximity effects and are unable to distinguish among isomers. The second level of estimation involves groups that provide better description of proximity effects and can differentiate among isomers. Hence, second level of estimation is intended to deal with polyfunctional, polar or nonpolar, and cyclic chemicals. The third level estimation includes groups that provide more structural information about molecular fragments of chemicals whose description is insufficient through the first- and second-order groups; hence, this level allows estimation of complex heterocyclic and polyfunctional acyclic chemicals. The MG method includes 220 first-order groups, 130 second-order groups, and 74 third-order groups to represent the molecular structure of the organic chemicals. The property prediction model to estimate the properties of organic chemicals employing MG method has the form[16]

$$f(X) = \sum_i N_i C_i + w \sum_j M_j D_j + z \sum_k E_k O_k \quad (1)$$

The function $f(X)$ is a function of property X, and it may contain additional adjustable model parameters (universal constants) depending on the property involved. In eq 1, C_i is the contribution of the first-order group of type i that occurs N_i times. D_j is the contribution of the second-order group of type j that occurs M_j times. E_k is the contribution of the third-order group of type k that has O_k occurrences in a component. Equation 1 is a general model for all the properties and the definition of $f(X)$ is specific for each property X. For determination of the contributions, C_i, D_j, and E_k, Marrero and Gani[16] suggested a three-step regression procedure.

- Step 1: In this step, the constants w and z are assigned zero values because only contributions of the first-order groups are estimated, that is, the first-order groups, C_i and the additional adjustable parameters of the model.

$$f(X) = \sum_i N_i C_i \quad (2)$$

- Step 2: In this step, the constants w and z are assigned unity and zero values, respectively, because only first and second-order groups are considered. The regression is performed (by keeping fixed the C_i and the adjustable parameters obtained from step 1) to determine the contributions of the second-order groups, D_j.

$$f(X) = \sum_i N_i C_i + \sum_j M_j D_j \quad (3)$$

- Step 3: In this step, both w and z are set to unity and regression is performed (by keeping fixed the obtained C_i, D_j, and the adjustable parameters obtained from steps 1 and 2) to determine the contributions of the third-order groups, E_k .

$$f(X) = \sum_i N_i C_i + \sum_j M_j D_j + \sum_k E_k O_k \quad (4)$$

In this way, the contributions of higher levels act as corrections to the approximations of the lower levels. Hukkerikar et al.[28] discussed a new approach for estimating the contributions, C_i, D_j, and E_k, based on the simultaneous regression method in which regression is performed by considering all of the terms of eq 1 to obtain contributions of first-, second-, and third-order groups in a single regression step.

Atom Connectivity Index (CI) Method. This method employs the following model for the estimation of properties of organic chemicals:[17]

$$f(X) = \sum_i a_i A_i + b({}^v\chi^0) + 2c({}^v\chi^1) + d \quad (5)$$

Where a_i is the contribution of the atom of type i that occurs A_i times in the molecular structure, ${}^v\chi^0$ is the zeroth-order (atom) valence connectivity index, ${}^v\chi^1$ is the first-order (bond) valence connectivity index, b and c are adjustable parameters, and d is a universal parameter. Please note that $f(X)$ of models in the MG method[16] and in the CI method[17] (i.e., left-hand side of eqs 1 and 5) has the same functional form for a particular pure component property X and the values of universal constants for the CI models are the same as those for the GC models.

Group-Contribution$^+$ (GC^+) Approach. For the purpose of creating missing groups and/or missing group contributions, the GC^+ approach is followed. The parameters, ${}^v\chi^0$ and ${}^v\chi^1$ for the groups as well as for the entire molecule are calculated using the rules described by Gani et al.[17] Once these indices are calculated, following CI model equations are applied to the missing groups to compute $f(X_m)$ and $f(X^*)$.

$$f(X_m) = \sum_i a_{m,i} A_{m,i} + b({}^v\chi^0)_m + 2c({}^v\chi^1)_m \quad (6)$$

$$f(X^*) = \left(\sum_m n_m f(X_m)\right) + d \quad (7)$$

Where m is the number of different missing groups and n_m indicates the number of times a missing group appears in the molecule. Finally, value of property X is estimated using the following equation of GC^+ method.

$$f(X) = \sum_i N_i C_i + f(X^*) + \sum_j M_j D_j + \sum_k E_k O_k \quad (8)$$

Database. For the estimation of property model parameters, large experimental data sets of organic chemicals of various classes (hydrocarbons, oxygenated components, nitrogenated components, poly functional components, etc.) from the

dx.doi.org/10.1021/ci300350r | J. Chem. Inf. Model. 2012, 52, 2823−2839

Table 2. Description of the Data Sets Used for Regression Purpose

(a) US EPA

class of chemicals	LC_{50} (FM)	LC_{50} (DM)	LD_{50}	Log W_s	BCF	PCO	PEL	GWP	ODP	AP
hydrocarbons	32	19	69	236	79	337	98	0	0	0
oxygenated	238	54	1382	1110	76	244	127	1	0	0
nitogenated	80	24	397	244	57	8	45	0	0	0
chlorinated	48	37	111	274	77	23	41	5	3	5
fluorinated	1	0	3	21	1	5	4	23	0	0
brominated	10	4	14	47	15	5	7	2	1	0
iodinated	1	0	5	17	0	0	1	0	0	0
phosphorus containing	0	0	5	0	0	0	0	0	0	0
sulfonated	9	8	24	19	5	0	15	0	0	0
silicon containing	0	0	1	2	0	0	0	0	0	0
multifunctional	390	174	3984	2711	352	17	87	20	24	5
total number of chemicals	809	320	5995	4681	662	639	425	51	28	10

(b) USEtox

class of chemicals	EUA_C	EUA_{NC}	ERA_C	ERA_{NC}	EFW_C	EFW_{NC}	ESW_C	ESW_{NC}	ENS_C	ENS_{NC}	EAS_C	EAS_{NC}
hydrocarbons	25	14	18	16	19	14	19	16	18	16	20	16
oxygenated	107	56	96	60	98	57	101	60	96	58	97	58
nitogenated	29	14	27	14	27	13	26	15	27	15	27	14
chlorinated	46	23	43	26	44	27	45	32	45	30	43	28
fluorinated	4	1	4	1	4	1	4	1	4	1	4	1
brominated	6	2	5	2	4	2	5	2	5	3	5	3
iodinated	0	0	0	0	0	0	0	0	0	0	0	0
phosphorus containing	0	0	0	0	0	0	0	0	0	0	0	0
sulfonated	3	1	3	1	3	1	3	1	3	1	3	1
silicon containing	0	0	0	0	0	0	0	0	0	0	0	0
multifunctional	236	230	274	229	273	230	274	233	262	238	271	231
total number of chemicals	456	341	470	349	472	345	477	360	460	362	470	352

database of US Environmental Protection Agency (EPA)[10] and from the database of USEtox[7] is used. The details of data set of each property in terms of number of organic chemicals belonging to various classes are given in Table 2a (data sets from US Environmental Protection Agency (EPA)) and in Table 2b (data sets from USEtox).

Parameter Estimation and Uncertainty Analysis (Maximum-Likelihood Estimation). The following discussion on parameter estimation and uncertainty analysis is based on the methodology discussed by Hukkerikar et al.[28] Let the property prediction model be represented by f and the model parameters (group/atom contributions, adjustable parameters, and universal parameter) by $\boldsymbol{P}$. The minimization of a cost function, $S(\boldsymbol{P})$, defined as the sum of the squares of the difference between the experimental value, X^{exp}, and evaluated property value, X^{pred}, provides the values of unknown parameters $\boldsymbol{P}^*$. This implies that $\boldsymbol{P}^*$ is a set of model parameter values obtained at the minimum value of the cost function value.

$$S(\boldsymbol{P}) = \min \sum_{j=1}^{N} (X_j^{exp} - X_j^{pred})^2 \tag{9}$$

The subscript j indicates the chemical evaluated, and N is the total number of chemicals included in the evaluation. After the estimation of the model parameters, uncertainty analysis is performed to quantify the model prediction errors. In this work, since the proposed models for environment-related properties are linear in nature, the following discussion is intended to provide information on linear least-squares theory. For linear least-squares, the covariance matrix of the estimated model parameters, COV($\boldsymbol{P}^*$), is given by,[29]

$$\mathrm{COV}(\boldsymbol{P}^*) = \frac{\mathrm{SSE}}{\nu}(\boldsymbol{A}^T\boldsymbol{A})^{-1} \tag{10}$$

Where, SSE is the sum of squared errors obtained from the least-squares parameter estimation method, ν is the degrees of freedom (that is, the total number of measurements, n, minus the number of unknown parameters, m). For the GC model with linear form of $f(X)$, $\boldsymbol{A}$ is the matrix containing frequencies (or occurrences) of groups used to represent the chemicals in the data set used for the regression. For the CI model with linear form of $f(X)$, $\boldsymbol{A}$ is the matrix containing frequencies of atoms and zeroth-order and first-order connectivity index for each chemical included in the data set. The covariance matrix computed using eq 10 is used for assessing the quality of the parameter estimation. The diagonal elements of this matrix are the variances of the errors of the parameter estimates and the off-diagonal elements are the covariances between the parameter estimation errors.

The confidence interval of the parameters, $\boldsymbol{P}^*$, at α_t significance level is given as,[29,30]

$$\boldsymbol{P}^*_{1-\alpha_t} = \boldsymbol{P}^* \pm \sqrt{\mathrm{diag}(\mathrm{COV}(\boldsymbol{P}^*))}\cdot t(\nu, \alpha_t/2) \tag{11}$$

In eq 11, $t(\nu, \alpha_t/2)$ is the t-distribution value corresponding to the $\alpha_t/2$ percentile (α_t is usually a value of 0.05) and diag(COV($\boldsymbol{P}^*$)) represents the diagonal elements of COV($\boldsymbol{P}^*$). The t-distribution value is obtained from the probability distribution function of students' t-distribution,[31] $P_r(t, \nu)$, and is given as,

$$P_r(t, \nu) = \sqrt{\nu}B\left(\frac{1}{2}, \frac{\nu}{2}\right)^{-1}\int_{-t}^{t}(1 + x^2/\nu)^{-1/2(\nu+1)}\,dx \tag{12}$$

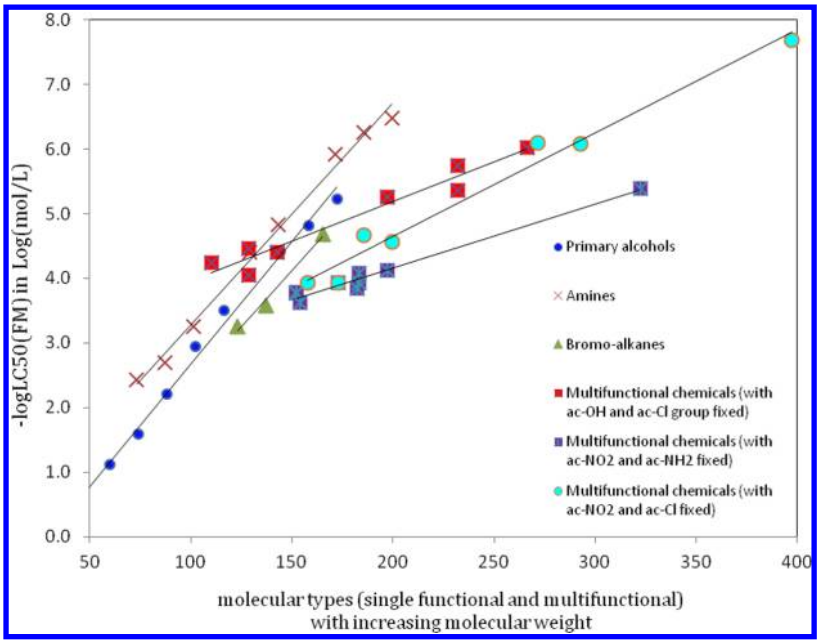


Figure 1. Plot of molecule types versus their experimental values of −Log LC_{50}(FM).

Where $x = \nu/(\nu + t^2)$ and $B\ (1/2, \nu/2)$ is the beta function. For 95% confidence interval calculation, the value of $P_r(t, \nu)$ is 0.95. The t-distribution value can also be obtained using the "tinv" function available in MatLab.

The confidence interval of the predicted property value, X^{pred}, at α_t significance level is given as

$$X^{pred}_{1-\alpha_t} = X^{pred} \pm \sqrt{\text{diag}(J(\boldsymbol{P}^*)\text{COV}(\boldsymbol{P}^*)J(\boldsymbol{P}^*)^T)} \cdot t(\nu, \alpha_t/2) \quad (13)$$

Where, the Jacobian matrix $J(\boldsymbol{P}^*)$ calculated using $\partial f/\partial \boldsymbol{P}^*$ represents the local sensitivity of the property model f to variations in the estimated parameter values $\boldsymbol{P}^*$. It is to be noted that the uncertainties of the property data arise mainly due to accuracy and precision of measurement instruments used, method of measurement, and purity of samples considered in the analysis, among others. The model prediction error reported as 95% confidence interval on calculated properties is a statistical concept associated with statistical framework used for parameter estimation. In this study, we use the maximum likelihood theory as summarized in eqs 9−13, which aims to propagate the residuals (that is, the difference between the property data and the calculated property values using the model) obtained after parameter estimation as first errors on model parameters (covariance matrix of estimated parameters) and then errors on the model predictions using linear error propagation method.[29] The model prediction error reported as 95% confidence interval is useful to assess the reliability of the prediction (when experimental data is available for the property). If the experimental value of the property is within the calculated confidence interval, then the property prediction method is verified as reliable. When experimental data is unavailable, the calculated confidence interval provides a measure of the likely prediction error (uncertainty) of the predicted property value. This information can be used in the design and analysis of sustainable processes to take into account the effect of uncertainties of predicted property values on the calculated impact that the processes would have on the environment (and hence on the decision of selection of sustainable process design).

Statistical Performance Indicators. The statistical significance of the developed correlations in this work is based on the following performance indicators.[28]

- Standard deviation (SD): This parameter measures the spread of the data about its mean value μ and is given by

$$\text{SD} = \sqrt{\sum_j (X_j^{exp} - X_j^{pred})^2/N} \quad (14)$$

- Average absolute error (AAE): This is the measure of deviation of predicted property values from the experimentally measured property values and is given by

$$\text{AAE} = \frac{1}{N}\sum_j |X_j^{exp} - X_j^{pred}| \quad (15)$$

- Average relative error (ARE): This provides an average of relative error calculated with respect to the experimentally measured property values and is given by

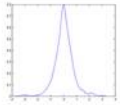
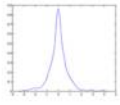
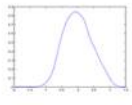
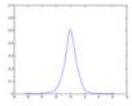
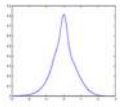
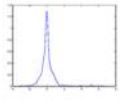
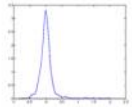
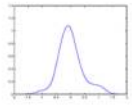
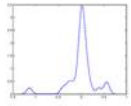
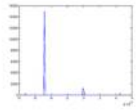
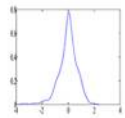
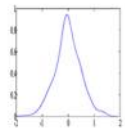
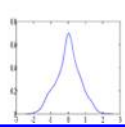
$$\text{ARE} = \frac{1}{N}\sum_j |(X_j^{exp} - X_j^{pred})/X_j^{exp}| \times 100 \quad (16)$$

- Coefficient of determination (R^2): This parameter provides information about the goodness of model fit. An R^2 close to 1.0 indicates that the experimental data used in the regression have been fitted to a good accuracy. It is calculated using,

$$R^2 = 1 - [\sum_j (X_j^{exp} - X_j^{pred})^2/\sum_j (X_j^{exp} - \mu)^2] \quad (17)$$

The indicators SD, AAE, ARE, and R^2 provide measures of quality (reliability) of property prediction models on a global basis. However, it is important that the information of uncertainties of estimated values also be made available to the

Table 3. Performance of MG Method Based Property Models Analysed Using Stepwise Regression Method

sl. no.	property	L.H.S. of MG method based property prediction model $f(X)$	MG group-contribution model $f(X)=\sum_i N_i C_i+\sum_j M_j D_j+\sum_k E_k O_k$									
			N	v	R^2	residual distribution plot	P_{rc} ($\pm$ 1%)	P_{rc} ($\pm$ 5%)	P_{rc} ($\pm$ 10%)	SD	AAE	ARE[a]
1	fathead minnow 96-h LC_{50} (LC_{50} (FM)) in mol/lit	$-\mathrm{LogLC_{50}(FM)}+\mathrm{FM_0}$	809	541	0.78		8.53	31.52	54.02	0.69	0.48	21.56
2	*daphnia magna* 48-h LC_{50} (LC_{50} (DM)) in mol/lit	$-\mathrm{LogLC_{50}(DM)}+\mathrm{DM_0}$	320	124	0.82		16.25	39.06	62.50	0.74	0.49	16.16
3	oral rat LD_{50} (LD_{50}) in mol/kg	$-\mathrm{LogLD50}-A_{LD50}-B_{LD50}\mathrm{MW}$	5995	5617	0.73		1.52	6.92	13.61	0.43	0.35	16.40
4	aqueous solubility ($LogW_S$) in mg/lit	$\mathrm{LogW_S}-A_{W_S}-B_{W_S}\mathrm{MW}$	4681	4311	0.78		3.12	14.36	28.63	0.99	0.73	----
5	bioconcentra -tion factor (BCF)	LogBCF	662	423	0.78		8.91	19.49	30.82	0.63	0.47	----
6	permissible exposure limit (OSHA-TWA) in mol/m^3	-LogPEL	425	239	0.74		16.71	39.53	60.24	0.78	0.44	12.61
7	photochemic al oxidation potential (PCO)	-LogPCO	639	488	0.83		6.42	16.9	26.30	0.22	0.13	8.37
8	global warming potential (GWP)	LogGWP	51	31	0.87		15.69	37.25	56.86	0.41	0.29	11.57
9	ozone depletion potential (ODP)	LogODP	28	12	0.89		17.86	21.4	28.5	0.30	0.16	----
10	acidification potential (ODP)	LogAP	10	1	1.0		100.0	--	--	3.4E-04	2.1E-4	----
11	emission to urban air (EUA_C) in cases/kg emitted (carcinogenic)	$-\mathrm{Log}(\mathrm{EUA_C})+A_{EUA_C}$	456	214	0.70		16.23	40.13	63.60	0.70	0.50	10.61
12	emission to urban air (EUA_{NC}) in cases/kg emitted (noncarcinogenic)	$-\mathrm{Log}(\mathrm{EUA_{NC}})+A_{EUA_{NC}}$	341	128	0.79		12.90	47.80	76.50	0.49	0.37	6.80
13	emission to continental rural air (ERA_C) in cases/kg emitted (carcinogenic)	$-\mathrm{Log}(\mathrm{ERA_C})+A_{ERA_C}$	470	229	0.75		15.74	39.15	64.89	0.67	0.51	8.88

dx.doi.org/10.1021/ci300350r | J. Chem. Inf. Model. 2012, 52, 2823−2839

Table 3. continued

sl. no.	property	L.H.S. of MG method based property prediction model $f(X)$	N	v	R^2	residual distribution plot	MG group-contribution model $f(X)=\sum_i N_i C_i+\sum_j M_j D_j+\sum_k E_k O_k$ P_{rc} (± 1%)	P_{rc} (± 5%)	P_{rc} (± 10%)	SD	AAE	ARE[a]
14	emission to continental rural air (ERA_{NC}) in cases/kg emitted (noncarcinogenic)	$-\mathrm{Log}(ERA_{NC})+A_{ERA_{NC}}$	349	134	0.80		13.18	46.13	75.07	0.55	0.42	7.25
15	emission to continental fresh water (EFW_C) in cases/kg emitted (carcinogenic)	$-\mathrm{Log}(EFW_C)+A_{EFW_C}$	472	230	0.75		13.77	31.77	60.16	0.67	0.52	11.26
16	emission to continental fresh water (EFW_{NC}) in cases/kg emitted (noncarcinogenic)	$-\mathrm{Log}(EFW_{NC})+A_{EFW_{NC}}$	345	131	0.83		13.33	44.63	67.82	0.52	0.40	8.15
17	emission to continental sea water (ESW_C) in cases/kg emitted (carcinogenic)	$-\mathrm{Log}(ESW_C)+A_{ESW_C}$	477	235	0.81		15.30	37.94	67.71	0.79	0.61	8.69
18	emission to continental sea water (ESW_{NC}) in cases/kg emitted (noncarcinogenic)	$-\mathrm{Log}(ESW_{NC})+A_{ESW_{NC}}$	360	146	0.85		14.16	46.38	72.22	0.69	0.51	8.41
19	emission to continental natural soil (ENS_C) in cases/kg emitted (carcinogenic)	$-\mathrm{Log}(ENS_C)+A_{ENS_C}$	472	231	0.76		13.98	39.61	63.55	0.72	0.55	9.28
20	emission to continental natural soil (ENS_{NC}) in cases/kg emitted (noncarcinogenic)	$-\mathrm{Log}(ENS_{NC})+A_{ENS_{NC}}$	362	148	0.79		14.91	48.06	71.27	0.61	0.46	7.27
21	emission to continental agri-cultural soil (EAS_C) in cases/kg emitted (carcinogenic)	$-\mathrm{Log}(EAS_C)+A_{EAS_C}$	470	228	0.75		13.61	41.06	65.74	0.67	0.51	9.36
22	emission to continental agri-cultural soil (EAS_{NC}) in cases/kg emitted (noncarcinogenic)	$-\mathrm{Log}(EAS_{NC})+A_{EAS_{NC}}$	352	138	0.80		16.19	48.29	74.71	0.54	0.41	6.92

[a]ARE is not defined for Log W_s, BCF, ODP, and AP since these properties have both positive and negative values.

user in order to provide confidence in the estimated property values and hence in the design of sustainable processes.

RESULTS

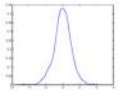
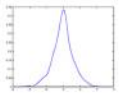
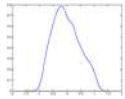
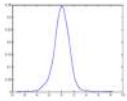
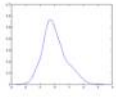
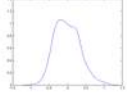
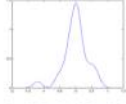
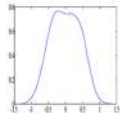
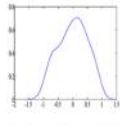
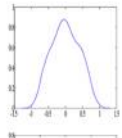
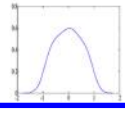
In this section, the selection of suitable property models for modeling environment-related properties and the performance statistics for the developed property models are discussed. The results are presented for the following models:

- MG method based property models analyzed using stepwise regression method
- MG method based property models analyzed using simultaneous regression method
- CI method based property models

Selection of Suitable Property Models for Environment-Related Properties. In this work, the basis for selecting an appropriate property model for the environment-related property has been the study of behavior of that property of certain class of chemicals with increasing molecular weight. This is illustrated for the case of LC_{50}(FM). Figure 1 shows plots of various molecules types (such as alcohols, amines, multifunctional chemicals, etc.) with increasing molecular weight versus their experimental values of $-\mathrm{Log}\ LC_{50}$(FM). It can be seen that these plots are linear in nature suggesting that the property LC_{50}(FM) can be modeled using a linear model of the form $-\mathrm{Log}\ LC_{50}(\mathrm{FM})$ + constant = $\sum_i N_i C_i + \sum_j M_j D_j + \sum_k E_k O_k$. Similar analyses have been performed (not shown in this paper) to obtain a suitable form of the property model for other

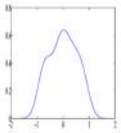
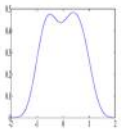
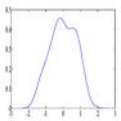
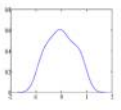
dx.doi.org/10.1021/ci300350r | J. Chem. Inf. Model. 2012, 52, 2823−2839

Table 4. Performance of CI Method Based Property Models

sl. no.	property	L.H.S. of CI method based property prediction model $f(X)$	N	v	R^2	residual distribution plot	atom connectivity index (CI) model $f(X)=\sum_i a_i A_i + b(^v\chi^0)+2c(^v\chi^1)+d$					
							P_{rc} (± 1%)	P_{rc} (± 5%)	P_{rc} (± 10%)	SD	AAE	ARE [a]
1	fathead minnow 96-h LC_{50} (LC_{50} (FM)) in mol/lit	$-\text{LogLC}_{50}\text{(FM)} + \text{FM}_0$	809	796	0.56		3.96	16.70	34.20	0.98	0.75	40.47
2	*daphnia magna* 48-h LC_{50} (LC_{50} (DM)) in mol/lit	$-\text{LogLC}_{50}\text{(DM)} + \text{DM}_0$	320	307	0.58		5.0	22.81	40.94	1.14	0.85	35.21
3	oral rat LD_{50} (LD_{50}) in mol/kg	$-\text{LogLD50} - A_{LD50} - B_{LD50}MW$	5662	5647	0.60		1.02	5.35	11.48	0.48	0.40	18.49
4	aqueous solubility (LogW_S) in mg/lit	$\log(W_S) - A_{W_S} - B_{W_S}MW$	4681	4676	0.62		2.22	9.98	19.80	1.29	0.98	----
5	bioconcentra -tion factor (BCF)	LogBCF	662	648	0.53		1.66	6.19	12.54	0.92	0.74	----
6	permissible exposure limit (PEL) in mol/m^3	-LogPEL	411	397	0.64		4.87	16.79	33.09	0.78	0.61	20.10
7	photochemic al oxidation potential (PCO)	-LogPCO	621	607	0.51		1.61	4.83	8.05	0.33	0.27	16.65
8	global warming potential (GWP)	LogGWP	51	37	0.83		9.80	31.37	50.98	0.48	0.36	15.52
9	ozone depletion potential (ODP)	LogODP	28	14	0.83		7.14	10.71	14.30	0.37	0.25	----
10	acidification potential (ODP)	LogAP	10	1	1.00		70.0	100.0	--	0.0014	802E-04	--
11	emission to urban air (EUA_C) in cases/kg emitted (carcinogenic)	$-\text{Log}(\text{EUA}_C) + A_{EUA_C}$	232	220	0.66		5.17	40.08	83.18	0.40	0.34	6.36
12	emission to urban air (EUA_{NC}) in cases/kg emitted (noncarcinogenic)	$-\text{Log}(\text{EUA}_{NC}) + A_{EUA_{NC}}$	259	247	0.66		7.72	40.92	69.11	0.49	0.41	7.50
13	emission to continental rural air (ERA_C) in cases/kg emitted (carcinogenic)	$-\text{Log}(\text{ERA}_C) + A_{ERA_C}$	226	214	0.79		11.94	49.55	88.05	0.39	0.32	5.43
14	emission to continental rural air (ERA_{NC}) in cases/kg emitted (noncarcinogenic)	$-\text{Log}(\text{ERA}_{NC}) + A_{ERA_{NC}}$	257	245	0.74		7.78	39.69	68.48	0.53	0.44	7.57

dx.doi.org/10.1021/ci300350r | J. Chem. Inf. Model. 2012, 52, 2823−2839

Table 4. continued

sl. no.	property	L.H.S. of CI method based property prediction model $f(X)$	N	ν	R^2	residual distribution plot	P_{rc} (±1%)	P_{rc} (±5%)	P_{rc} (±10%)	SD	AAE	ARE[a]
		atom connectivity index (CI) model $f(X)=\sum_i a_i A_i + b({}^{v}\chi^{0}) + 2c({}^{v}\chi^{1}) + d$										
15	emission to continental fresh water (EFW_C) in cases/kg emitted (carcinogenic)	$-\mathrm{Log}(EFW_C)+A_{EFW_C}$	286	274	0.65		7.34	36.36	61.53	0.52	0.44	8.51
16	emission to continental fresh water (EFW_{NC}) in cases/kg emitted (noncarcinogenic)	$-\mathrm{Log}(EFW_{NC})+A_{EFW_{NC}}$	259	247	0.70		9.26	33.59	60.61	0.54	0.44	9.02
17	emission to continental sea water (ESW_C) in cases/kg emitted (carcinogenic)	$-\mathrm{Log}(ESW_C)+A_{ESW_C}$	286	274	0.78		4.89	35.31	75.17	0.62	0.54	7.20
18	emission to continental sea water (ESW_{NC}) in cases/kg emitted (noncarcinogenic)	$-\mathrm{Log}(ESW_{NC})+A_{ESW_{NC}}$	291	279	0.77		5.84	32.30	65.63	0.72	0.60	8.76
19	emission to continental natural soil (ENS_C) in cases/kg emitted (carcinogenic)	$-\mathrm{Log}(ENS_C)+A_{ENS_C}$	285	273	0.61		6.66	38.59	74.38	0.52	0.44	6.89
20	emission to continental natural soil (ENS_{NC}) in cases/kg emitted (noncarcinogenic)	$-\mathrm{Log}(ENS_{NC})+A_{ENS_{NC}}$	247	235	0.70		9.31	39.67	72.06	0.53	0.45	7.08
21	emission to continental agricultural soil (EAS_C) in cases/kg emitted (carcinogenic)	$-\mathrm{Log}(EAS_C)+A_{EAS_C}$	240	228	0.68		8.33	42.50	88.33	0.42	0.36	5.76
22	emission to continental agricultural soil (EAS_{NC}) in cases/kg emitted (noncarcinogenic)	$-\mathrm{Log}(EAS_{NC})+A_{EAS_{NC}}$	247	235	0.70		8.50	40.89	74.08	0.49	0.42	6.94

[a]ARE is not defined for Log W_s, BCF, ODP, and AP since these properties have both positive and negative values.

environment-related properties with the objective of providing an accurate and reliable property estimation of environment-related properties.

Model Performance. The model performance statistics for property models analyzed using the stepwise regression method are provided in Table 3. The model performance statistics for properties analyzed using the simultaneous regression method are given in Table S1 in the Supporting Information. In Table 3, N is the number of experimental data points considered in the regression and ν is the degrees of freedom and is obtained by subtracting the number of estimated model parameters from N. $P_{rc(\pm 1\%)}$, $P_{rc(\pm 5\%)}$, and $P_{rc(\pm 10\%)}$ represents the percentage of the experimental data-points (N) found within ±1%, ± 5%, and ±10% relative error range, respectively. For property models analyzed using stepwise regression method, the results for R^2, SD, AAE, and ARE have been obtained after third-level estimation; hence, they represent the global results of the three sequential approximations. The residuals ($X^{exp} - X^{pred}$) for data points considered in the regression are plotted in the form of residual distribution plots and are included in Table 3 and Table S1 (Supporting Information). For most of the property models (except for ozone depletion potential and acidification potential), the residuals followed a normal distribution curve suggesting that the assumption of normal distribution of random errors is valid behind the followed approach. The user of a particular property model can decide the selection of stepwise or simultaneous method based on the performance statistics for that property given in Table 3 and Table S1. The model performance statistics for property models analyzed using the CI method are provided in Table 4. These CI models have been employed together with the GC method for creating the missing groups and predicting their contributions through the regressed contributions of connectivity indices as suggested by Gani et al.[17] This feature makes it possible to predict environment-related properties of organic chemicals for which neither experimental data nor the GC-property model parameters are available. The property models developed based on the CI method (see Table 4) have reasonable model performance statistics. High accuracy

dx.doi.org/10.1021/ci300350r | J. Chem. Inf. Model. 2012, 52, 2823−2839

Table 5. Performance of Model for (a) Oral Rat LD_{50}, (b) Fathead Minnow 96-h LC_{50}, (c) Emission to Urban Air (Carcinogenic), and (d) Emission to Urban Air (Non-Carcinogenic) Based on Different Combinations of Training Sets and Test Sets

(a) Oral Rat LD_{50}

data sets used for		model performance statistics for training set				model performance statistics for test set using the parameters estimated from regression of the training set			model performance statistics for test set using the parameters estimated from regression of the whole data set (containing 5995 data-points)		
training purpose	testing purpose	MSECV	SD Log (mol/kg)	AAE Log (mol/kg)	ARE %	SD Log (mol/kg)	AAE Log (mol/kg)	ARE %	SD Log (mol/kg)	AAE Log (mol/kg)	ARE %
A, B, C, D	E	0.1812	0.4257	0.3479	15.97	0.4628	0.3732	17.32	0.4220	0.3424	15.91
A, B, C, E	D	0.1796	0.4238	0.3456	15.90	0.4755	0.3839	17.72	0.4287	0.3506	16.21
A, B, D, E	C	0.1805	0.4248	0.3462	15.97	0.4754	0.3823	17.23	0.4251	0.3500	15.90
A, C, D, E	B	0.1788	0.4229	0.3449	15.89	0.4677	0.3813	17.20	0.4338	0.3536	16.00
B, C, D, E	A	0.1794	0.4236	0.3455	15.86	0.4694	0.3848	17.93	0.4302	0.3532	16.46
average performance		0.1799	0.4241	0.3460	15.91	0.4701	0.3811	17.48	0.4279	0.3499	16.09

(b) Fathead Minnow 96-h LC_{50}

data sets used for		model performance statistics for training set				model performance statistics for test set using the parameters estimated from regression of the training set			model performance statistics for test set using the parameters estimated from regression of the whole data set (containing 809 data-points)		
training purpose	testing purpose	MSECV	SD Log (mol/lit)	AAE Log (mol/lit)	ARE %	SD Log (mol/lit)	AAE Log (mol/lit)	ARE %	SD Log (mol/lit)	AAE Log (mol/lit)	ARE %
A, B, C, D	E	0.3400	0.5831	0.4015	19.68	1.3753	0.8615	27.04	0.6732	0.4786	15.68
A, B, C, E	D	0.3339	0.5778	0.3991	19.10	1.3944	0.9325	28.91	0.6854	0.4778	17.65
A, B, D, E	C	0.3624	0.6020	0.4237	14.99	1.3517	0.9127	47.19	0.6581	0.4802	33.19
A, C, D, E	B	0.3645	0.6037	0.4201	20.39	1.4857	0.9072	28.26	0.6399	0.4654	14.62
B, C, D, E	A	0.3453	0.5876	0.4142	17.62	1.5178	0.9710	35.12	0.6722	0.4831	21.19
average performance		0.3492	0.5908	0.4117	18.35	1.4249	0.9169	33.30	0.6657	0.4770	20.47

(c) Emission to Urban Air (Carcinogenic)

data sets used for		model performance statistics for training set				model performance statistics for test set using the parameters estimated from regression of the training set			model performance statistics for test set using parameters estimated from regression of the whole data set (containing 456 data-points)		
training purpose	testing purpose	MSECV	SD cases/kg emitted	AAE cases/kg emitted	ARE %	SD cases/kg emitted	AAE cases/kg emitted	ARE %	SD cases/kg emitted	AAE cases/kg emitted	ARE %
A, B, C, D	E	0.2050	0.4528	0.3024	5.86	1.8293	1.1974	27.82	0.5386	0.4135	9.16
A, B, C, E	D	0.2206	0.4697	0.3252	6.55	1.4664	0.9823	17.63	0.4766	0.3534	6.60
A, B, D, E	C	0.1675	0.4093	0.2871	5.52	1.6849	1.2268	25.79	0.6160	0.4165	9.52
A, C, D, E	B	0.2111	0.4595	0.3052	6.42	2.0921	1.3187	24.12	0.4854	0.3480	6.33
B, C, D, E	A	0.2182	0.4671	0.3267	6.63	1.3597	1.0115	19.55	0.4572	0.3478	6.55
average performance		0.2045	0.4517	0.3093	6.19	1.6865	1.1473	22.98	0.4713	0.3479	6.44

(d) Emission to Urban Air (Noncarcinogenic)

data sets used for		model performance statistics for training set				model performance statistics for test set using the parameters estimated from regression of the training set			model performance statistics for test set using parameters estimated from regression of the whole data set (containing 341 data-points)		
training purpose	testing purpose	MSECV	SD cases/kg emitted	AAE cases/kg emitted	ARE %	SD cases/kg emitted	AAE cases/kg emitted	ARE %	SD Cases/kg emitted	AAE Cases/kg emitted	ARE %
A, B, C, D	E	0.0655	0.2560	0.1716	3.14	2.3797	1.4854	26.38	0.3872	0.2801	5.03
A, B, C, E	D	0.0997	0.3157	0.2217	4.12	1.7248	1.0980	19.22	0.3693	0.2708	4.89
A, B, D, E	C	0.0831	0.2882	0.1894	3.50	2.4662	1.5963	30.72	0.3684	0.2615	5.01
A, C, D, E	B	0.0846	0.2909	0.1936	3.51	3.0908	1.6401	29.28	0.3980	0.3007	5.62
B, C, D, E	A	0.1097	0.3313	0.2289	4.23	8.7061	3.9458	80.85	0.3096	0.2166	3.80
average performance		0.0885	0.2964	0.2010	3.69	3.6735	1.9531	37.28	0.3665	0.2659	4.87

in the prediction of environment-related properties cannot be expected from this model, since only a few parameters are involved to represent a large data set of chemicals. Greater accuracy can be obtained by adding higher-order connectivity indices. However, the main objective of analyzing CI models in this work is to obtain the missing group contributions, for which only the first two connectivity indices should be sufficient.[17] Hukkerikar et al.[28] discussed the effect of quantity of experimental data on the quality of parameter estimation and illustrated that by including all of the available experimental data of the property in the regression it is possible to improve the predictive capability and application range of the property model. Therefore, in this work, we have considered all of the available experimental data of properties of chemicals for modeling environment-related properties. To illustrate this point, we have considered here an analysis of property model for oral rat LD_{50}, fathead minnow 96-h LC_{50}, and emission to continental urban air (carcinogenic (EUA_C) and noncarcinogenic (EUA_{NC})). The

dx.doi.org/10.1021/ci300350r | J. Chem. Inf. Model. 2012, 52, 2823−2839

whole experimental data sets of these properties (5995 data-points for oral rat LD_{50}, 809 data-points for fathead minnow 96-h LC_{50}, 456 data-points for emission to continental urban air (carcinogenic (EUA_C), and 341 data-points for emission to continental urban air (noncarcinogenic (EUA_{NC}))) is divided randomly in five subsets (A, B, C, D, and E) of equal size. The property model is trained on four subsets (using simultaneous regression method) and one subset is used for testing purpose. This procedure is repeated five times so that all subsets are used for testing purposes. The results in terms of SD, AAE, and ARE for training sets and for test sets is presented in Table 5a for oral rat LD_{50}, Table 5b for fathead minnow 96-h LC_{50}, Table 5c for emission to continental urban air (carcinogenic (EUA_C), and in Table 5d for emission to continental urban air (noncarcinogenic (EUA_{NC})). The MSECV, which is mean squared error of cross-validation,[32] calculated using eq 18 is also given in Table 5.

$$\mathrm{MSECV} = \frac{1}{N_L}\sum_{k=1}^{K}\sum_{j\in L_k}(X_j^{\exp} - X_j^{\mathrm{pred}})^2 \qquad (18)$$

Where, N_L is the number of data points in the training set, K = number of subsets (5 in this analysis), and L_k is the number of data-points in the subsets.

From Table 5a, comparison of the model performance for training sets and test sets show that the predictive capability of the model for oral rat LD_{50} is fairly good. This is mainly due to the large amount of available experimental data of oral rat LD_{50} for the training purpose. For test sets, if we compare the SD, AAE, and ARE values calculated using the parameters obtained by regressing training set with those that are calculated using the parameters obtained by regression of the whole data set, we find that better model performance statistics (lower SD, lower AAE, and lower ARE) is obtained when we use model parameters that are estimated using all of the experimental data-points in the regression.

For fathead minnow 96-h LC_{50}, emission to continental urban air (carcinogenic (EUA_C)), and emission to continental urban air (noncarcinogenic (EUA_{NC})), it can be seen from Table 5b−d that the model performance for test sets is poor as compared to those of training sets, and this is due to the small amount of available experimental data of these properties for the training purpose. For these properties, it can be observed that the SD, AAE, and ARE values for test sets calculated using the model parameters as obtained by regression of the whole data set are much better than those that are calculated using the parameters estimated using the training set indicating the importance of considering all of the available experimental data-points for the regression purpose. To sum up, this analysis shows both the robustness of the approach and the predictive capability of the developed models for estimating environmental related properties.

Marrero and Gani[22] reported SD, AAE, and R^2 values for the GC model for Log W_s as 0.55, 0.46, and 0.93, respectively. In their analysis, the number of estimated model parameters (groups) are 155 first-order groups, 99 second-order groups, and 48 third-order groups (that is, total 302 groups estimated out of 424 groups). Referring to Table 3, it can be seen that the property model for Log W_s has SD, AAE, and R^2 values of 0.99, 0.73, and 0.78 respectively. In this work, the number of estimated groups is 197 first-order groups, 124 second-order groups, and 57 third-order groups (total 378 groups estimated out of 424 groups). It is to be noted that in the present work, a much larger data set (4681 data points as compared to 2087 data points used by Marrero and Gani[22]) of Log W_s comprising complex and polyfunctional environment-related chemicals is used in the regression, which makes it possible to estimate larger number of model parameters thereby contributing to improved application range of the property model for Log W_s. A similar note can be made for the case of property model for LC_{50}(FM). The developed property model for LC_{50}(FM) has SD, AAE, and R^2 values of 0.69, 0.48, and 0.78, respectively. Martin and Young[20] reported SD and R^2 values for their GC model for LC_{50}(FM) as 0.37 and 0.91, respectively. The use of the large data set for LC_{50}(FM) allows estimation of a large number of model parameters which in turn allows one to estimate LC_{50}(FM) for a wide range of organic chemicals. For the property LC_{50}(DM), the model performance statistics are similar to that of LC_{50}(FM) model. The developed property model for LD_{50} (using a data set of 5995 chemicals) has reasonably good performance statistics with SD, AAE, and R^2 values as 0.43, 0.35, and 0.73, respectively. Several estimation methods based on the QSAR approach have been reported in the literature that uses other properties such as LC_{50}(DM) as an input to their estimation method to estimate LD_{50}. Also, these methods have been developed based on relatively smaller data sets (with few hundreds of chemicals in the data set) of chemicals. The application of such methods is restricted by the availability of the experimental data of the needed input properties for their estimation. A similar issue is associated with the estimation methods for BCF requiring additional inputs such as the octanol/water partition coefficient. In this work, the developed property model for BCF has SD, AAE, and R^2 values of 0.63, 0.47, and 0.78, respectively. Note that the developed property models for LD_{50} and for BCF only require the molecular structure of the chemical for the property estimation. For properties GWP, ODP, and AP, the number of experimental data points used in the regression are smaller as compared to other properties analyzed in this work. However, it can be noted that these properties belong to a particular class of chemicals (for example, global warming potential and ozone depletion potential properties involve halogenated chemicals, acidification potential property involves nitrogenated chemicals); hence, even though the experimental data sets are smaller, the models for these properties are able to provide estimation of these properties with good accuracy. The model performance statistics for the remaining properties namely, EUA_C, EUA_{NC}, ERA_C, ERA_{NC}, EFW_C, EFW_{NC}, ESW_C, ESW_{NC}, ENS_C, ENS_{NC}, EAS_C, and EAS_{NC}, show that the experimental data have been fitted to a good degree of accuracy. The estimation of these properties is based exclusively on the molecular structure of the chemical and allows the user to calculate human toxicity potential (HTP)[8] (which is needed to perform life cycle impact assessment of the product) thus increasing the application range of the USEtox model[7] to a wide range of chemicals.

The variables FM_0, DM_0, A_{LogWs}, B_{LogWs} A_{LD50}, B_{LD50}, A_{EUAC}, A_{EUANC}, A_{ERAC}, A_{ERANC}, A_{EFWC}, A_{EFWNC}, A_{ESWC}, A_{ESWNC}, A_{ENSC}, A_{ENSNC}, A_{EASC}, A_{EASNC} as defined in the functional forms, $f(X)$, given in Tables 3 and 4 are additional adjustable parameters of property prediction models. The values of these parameters are listed in Table 6. The total list of groups and their contributions C_j, D_j, and E_k for the 22 environment-related properties analyzed in this work are given in the Supporting Information (see Tables S2−S4 for MG method based models analyzed using stepwise regression method and Tables S5−S7 for MG method based models analyzed using simultaneous regression method). The list of atoms, their contributions a_i, adjustable model parameters (b and c), and the universal parameter d for CI method based

Table 6. Values of Universal Constants (Additional Adjustable Parameters)[a]

universal constants	value (stepwise method)	value (simultaneous method)
FM_0	2.1949	2.1841
DM_0	2.9717	3.5907
A_{LogWs}	4.5484	4.3098
B_{LogWs}	0.3411	0.3404
A_{LD50}	1.9372	1.9372
B_{LD50}	0.0016	0.0016
A_{EUAC}	5.2801	5.22536
A_{EUANC}	6.8181	7.06605
A_{ERAC}	6.5561	6.68611
A_{ERANC}	7.5541	9.53269
A_{EFWC}	5.6726	5.0706
A_{EFWNC}	6.4429	7.33378
A_{ESWC}	8.3962	9.33319
A_{ESWNC}	8.6360	10.0724
A_{ENSC}	6.4837	5.93334
A_{ENSNC}	7.0265	6.4159
A_{EASC}	6.2913	5.48504
A_{EASNC}	6.9723	6.06003

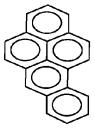
[a]Values of universal constants for the CI models are the same as those based on the stepwise method.

property prediction models are given in the Supporting Information (see Table S8). The covariance matrix computed using eq 10 for each property prediction model analyzed using the MG method (for models with stepwise regression method and simultaneous regression method) and using the CI method is available upon request from the authors. The developed models for environment-related properties have been implemented in ProPred, a property estimation toolbox of ICAS (Integrated Computer Aided System[33]) software developed by CAPEC, DTU.

Application of the Developed Property Models for the Estimation of Environment-Related Properties. The application of the developed property models to estimate environment-related properties and to quantify the uncertainties of the estimated property values is illustrated by considering predictions of Log W_s (using model parameters obtained from simultaneous regression method) for the chemical, benzo[*a*]-pyrene, (CAS no. 50-32-8) which is a polycyclic aromatic hydrocarbon and is highly carcinogenic. The experimentally measured value of Log W_s (mg/L) for benzo[*a*]pyrene is −2.79. Table 7 provides information of first-order, second-order, and third-order groups used to represent benzo[*a*]pyrene, their frequency (that is, occurrences in the structure), and the contributions for each group (Log W_{s1i}, Log W_{s2j}, and Log W_{s3k}) taken from Tables S5−S7 given in the Supporting Information. Using this information and the universal constants of the property model for Log W_s, the value of Log W_s for benzo[*a*]pyrene is estimated as −2.64 (with absolute error = |−2.79 − (−2.64)| = 0.15).

To illustrate the application of the GC$^+$ approach, let us assume that the contribution of aC fused with aromatic ring group is not available. In this case, the model user cannot apply the GC method for predicting Log W_s of benzo[*a*]pyrene. However, using the GC$^+$ approach, it is possible to predict the missing contribution of aC fused with aromatic ring group which will allow the model user to predict Log W_s of benzo[*a*]pyrene. The zeroth-order (atom) connectivity index, $^v\chi^0$, and the first-order (bond) connectivity index, $^v\chi^1$, for aC fused with aromatic ring group as calculated using the rules given by Gani et al.[17] are 2.1380 and 1.7412, respectively. The CI model constants *b*, *c*, and *d* for Log W_s are −0.2601, −0.0376, and 0, respectively. Using eq 6, we predict the contribution of aC fused with aromatic ring group as −3.5296. This predicted contribution can now be substituted in eq 8 together with contributions of other groups listed in Table 7 to estimate the value of Log W_s, and this value is −1.7378 (the absolute error is therefore, 1.05).

Table 7. Estimation of Log W_s of Benzo[*a*]pyrene

Benzo[a]pyrene	molecular structure	
CAS No. 50-32-8		
molecular formula: $C_{20}H_{12}$		
first-order groups	**occurrences**	**contribution**
aCH	12	-4.5565
aC fused with aromatic ring	8	-4.7557
second-order groups	**occurrences**	**contribution**
No second-order groups are involved		
third-order groups	**occurrences**	**contribution**
AROM.FUSED[2]	1	-0.0759
AROM.FUSED[3]	1	-0.1255
AROM.FUSED[4p]	2	0.0500

$$\mathrm{LogW_s} = \mathrm{A_{LogWs}} + (\mathrm{B_{LogWs}} \cdot \mathrm{MW}) + \sum_i N_i C_i + w \sum_j M_j D_j + z \sum_k E_k O_k = -2.64$$

dx.doi.org/10.1021/ci300350r | J. Chem. Inf. Model. 2012, 52, 2823−2839

Table 8. Covariance Matrix COV(P^*) with Dimensions 7 × 7

	A_{LogWs}	B_{LogWs}	aCH	aC	arom.fused[2]	arom.fused[3]	arom.fused[4p]
A_{LogWs}	0.0154						
B_{LogWs}	-1.28×10^{-7}	4.97×10^{-7}					
aCH	−0.0025	-8.1×10^{-6}	6.71×10^{-4}				
aC	−0.002	-4.7×10^{-6}	-3.7×10^{-4}	0.0048			
arom.fused[2]	7.7×10^{-5}	-4.3×10^{-6}	-5.9×10^{-4}	−0.0047	0.0113		
arom.fused[3]	−0.0013	-2.4×10^{-6}	-8.9×10^{-6}	−0.0084	0.0111	0.0375	
arom.fused[4p]	-4.5×10^{-4}	1.6×10^{-6}	8.25×10^{-6}	−0.0092	0.009	0.0136	0.0283

Table 9. Local Sensitivity $J(P^*)$ with Dimensions (1 × 7) of Log W_s Model with Respect to the Model Parameters

δLog $W_s/\delta A_{LogWs}$	δLog W_s /δB_{LogWs}	δLog W_s //δaCH	δLogW_s //δaC	δLog W_s /δ arom.fused[2]	δLog W_s /δ arom.fused[3]	δLog W_s /δ arom.fused[4p]
1.0	252.31	12	8	1	1	2

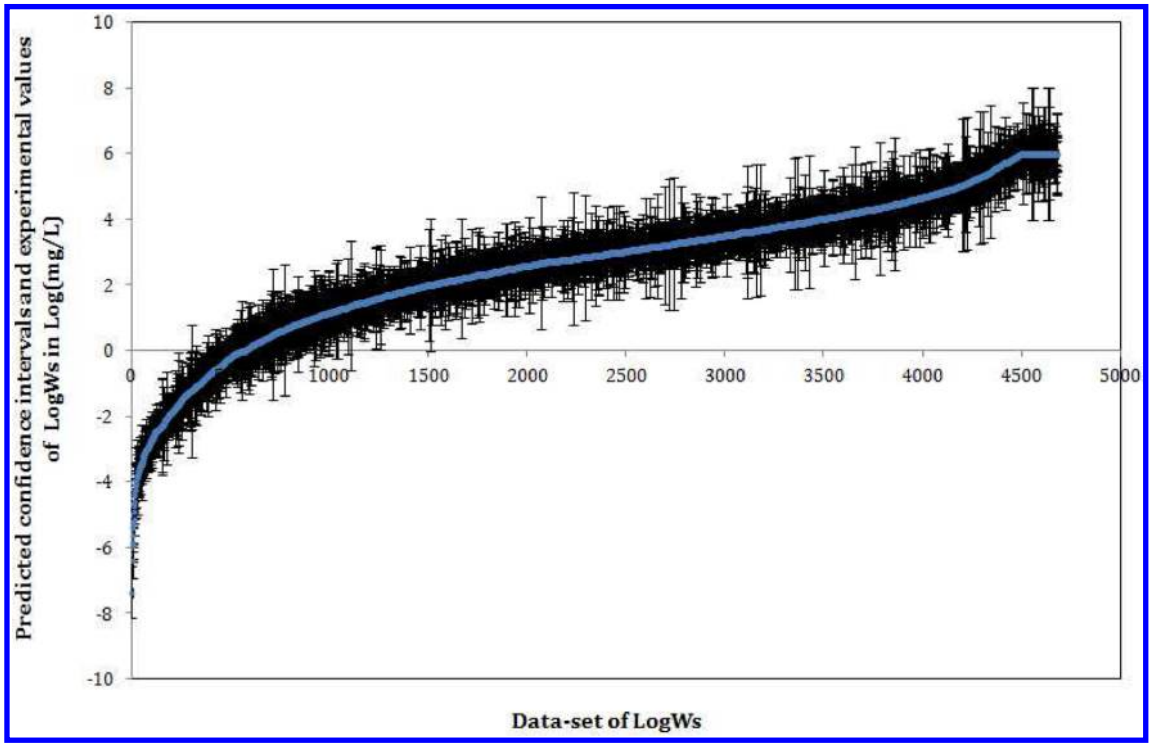


Figure 2. Experimental values of Log W_s and the calculated 95% confidence intervals versus data set of Log W_s.

As a next step, the uncertainty of the estimated Log W_s is quantified. For this purpose, information of covariance COV(P^*) of the involved groups and the universal constants A_{LogWs} and B_{LogWs} and also the local sensitivity $J(P^*)$ of the Log W_s model is needed. The covariance of the involved groups (as listed in Table 8) and universal constants A_{LogWs} and B_{LogWs} was noted from the overall covariance matrix for all the groups of the Log W_s model analyzed using simultaneous regression method. In Table 8, only lower triangular elements are shown since the upper triangular matrix elements are identical to the lower ones. Table 9 lists the local sensitivity of the Log W_s model with respect to the model parameters (for contributions listed in Table 7 and universal constants A_{LogWs} and B_{LogWs}).

To calculate the confidence intervals of estimated property values, say the 95% confidence intervals of the estimated Log W_s value, the covariance matrix COV(P^*) given in Table 8 and the local sensitivity $J(P^*)$ given in Table 9 are substituted in eq 13. For 95% confidence interval calculation, the t-distribution value corresponding to 0.05/2 percentile (i.e., $\alpha_t/2$ percentile) and with 4311 degrees of freedom (taken from Table 3) is obtained by solving eq 12 for t and this value is 1.9604. The predicted value of the Log W_s is −2.64 (see Table 7). The calculated 95% confidence intervals of the estimated Log W_s value is therefore

$$\underbrace{\mathrm{LogW_s}^{pred}_{(1-0.05)}}_{} = \underbrace{\mathrm{LogW_s}^{pred}}_{-2.64} \pm \underbrace{\sqrt{\mathrm{diag}\left(J(\boldsymbol{P}^*)COV(\boldsymbol{P}^*)J(\boldsymbol{P}^*)^T\right)}}_{0.2134} \cdot \underbrace{t(\nu,\alpha_t/2)}_{1.9604} = -2.64 \pm 0.41$$

It can be observed that the experimental value of the Log W_s (−2.79) lies within the predicted confidence intervals indicating reliability of the developed model for estimating property values of Log W_s and uncertainties of the estimated values. This, of course, can only be checked when experimental data is available. This is further illustrated in Figure 2 by plotting the experimental values of Log W_s and the calculated 95% confidence intervals (shown as vertical bars) for the entire experimental data set of Log W_s used for the regression purpose. About 42% of the experimental values in the data set (with 4681 data points) of Log W_s falls within the confidence intervals calculated at 95% confidence level. For the case where no experimental data is available, the calculated confidence intervals provide a measure of

dx.doi.org/10.1021/ci300350r | J. Chem. Inf. Model. 2012, 52, 2823−2839

the likely prediction error (that is, uncertainty) of the predicted property value. We have considered here the calculation of confidence intervals of the estimated property values using models analyzed by simultaneous regression method in order to simplify the illustration of the application of the developed property models, since there will be a single covariance matrix containing covariance of all the listed groups and universal parameters. The approach discussed in this section is the same for the case of property models analyzed using the stepwise regression method. In the case of stepwise regression method, there will be a covariance matrix for each type of the groups, i.e., first-order, second-order, and third-order and, hence, quantification of uncertainty in the predicted property value is to be performed (using these covariance matrices) for each step (that is step 1, step 2, and step 3) of property estimation.

CONCLUSIONS

Property models for environment-related properties based on the GC$^+$ approach have been developed with the objective of providing reliable estimation of these properties together with the uncertainties of the estimated values for their use in the synthesis, design, and analysis of sustainable processes. The estimation of environment-related properties using these models requires only the molecular structure of the organic chemicals. Large experimental data sets of environment-related properties taken from the database of US Environmental Protection Agency (EPA) and from the database of USEtox are used for the regression purpose in order to achieve good model performance and large application range of the property models. In total, 22 environment-related properties of organic chemicals have been modeled and analyzed. The use of the developed property models to estimate environment-related properties and the uncertainties of the estimated property values is illustrated through an application example. The models for some of the properties analyzed in this work have been implemented in ProPred, a property estimation toolbox of ICAS (Integrated Computer Aided System) software. The developed property models provide reliable estimates of environment-related properties needed to perform design and analysis of sustainable processes and allow one to evaluate the effect of uncertainties of estimated property values on the calculated potential impact that the processes would have on the environment giving useful insights into quality and reliability of the design of sustainable processes. Our current and future work is focused on quantification of the effect of uncertainties of estimated properties on the design of sustainable processes.

ASSOCIATED CONTENT

Supporting Information

Table S1: performance of MG method based property models analyzed using the simultaneous regression method. Table S2: MG method based property models analyzed using the stepwise regression method: first-order groups and their contributions for the environment-related properties. Table S3: MG method based property models analyzed using the stepwise regression method: second-order groups and their contributions for the environment-related properties. Table S4: MG method based property models analyzed using the stepwise regression method: third-order groups and their contributions for the environment-related properties. Table S5: MG method based property models analyzed using the simultaneous regression method: first-order groups and their contributions for the environment-related properties. Table S6: MG method based property models analyzed using the simultaneous regression method: second-order groups and their contributions for the environment-related properties. Table S7: MG method based property models analyzed using the simultaneous regression method: third-order groups and their contributions for the environment-related properties. Table S8: CI method based property models: atom contributions and model constants for the environment-related properties. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail address: rag@kt.dtu.dk. Phone: +45-45252882. Fax: +45-45932906.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge U.S. EPA, National Risk Management Research Laboratory, Cincinnati, Ohio, for providing experimental data for this work. This research work is a part of the collaboration between Technical University of Denmark, Denmark, and Alfa Laval Copenhagen A/S, Denmark. The research leading to these results has received funding from the European Union Seventh Framework Programme [FP7-MC-ITN] under grant agreement no. 238013.

ABBREVIATIONS

AAE, average absolute error; a_i, contribution of atom of type-i; A_i, occurrence of atom of type-i; ARE, average relative error [%]; b, adjustable parameter of eq 5; B (1/2, ν/2), beta function; c, adjustable parameter of eq 5; CI, atom connectivity index; C_i, contribution of first-order group of type-i; COV($\boldsymbol{P}$*), covariance matrix; d, universal parameter of eq 5; D_j, contribution of second-order group of type-j; E_k, contribution of third-order group of type-k; $f(X)$, function of property X; GC, group-contribution; GC$^+$, group-contribution$^+$; $J(\boldsymbol{P}^*)$, local sensitivity of the model to variations in estimated model parameters; MG, Marrero and Gani; M_j, occurrence of second-order group of type-j; MSECV, mean squared error of cross-validation; N, number of experimental data-points used in the regression; N_i, occurrence of first-order group of type-i; O_k, occurrence of third-order group of type-k; $\boldsymbol{P}$, model parameters; $\boldsymbol{P}^*$, estimated values of model parameters; $P_r(t, \nu)$, students t-distribution function; P_{rc}, percentage of the experimental data points [%]; R^2, coefficient of determination; $S(\boldsymbol{P})$, cost function; SD, standard deviation; SSE, minimum sum of squared errors; $t(\nu, \alpha_t/2)$, t-distribution value corresponding to the $\alpha_t/2$ percentile; X^{exp}, experimental property value; X^{pred}, predicted property value

Greek Symbols

$^v\chi^0$, zeroth-order (atom) connectivity index; $^v\chi^1$, first-order (bond) connectivity index; ν, degrees of freedom

REFERENCES

(1) Carvalho, A.; Gani, R.; Matos, H. Design of sustainable processes: Systematic generation and evaluation of alternatives. *Comput.-Aided Chem. Eng.* **2006**, *21*, 817−822.

(2) Cabezas, H.; Bare, J. C.; Mallick, S. K. Pollution prevention with chemical process simulators: The generalized waste reduction (WAR) algorithm − full version. *Comput. Chem. Eng.* **1999**, *23* (4−5), 623−634.

(3) Young, D. M.; Cabezas, H. Designing sustainable processes with simulation: The waste reduction (WAR) algorithm. *Comput. Chem. Eng.* **1999**, *23* (10), 1477−1491.

(4) Young, D. M.; Sharp, R.; Cabezas, H. The waste reduction (WAR) algorithm: Environmental impacts, energy consumption, and engineering economics. *Waste Manage.* **2000**, *20*, 605−615.
(5) Jensen, N.; Coll, N.; Gani, R. An integrated computer-aided system for generation and evaluation of sustainable process alternatives. *Clean Technol. Environ. Policy.* **2003**, *5*, 209−225.
(6) Carvalho, A.; Gani, R.; Matos, H. Design of sustainable chemical processes: Systematic retrofit analysis generation and evaluation of alternatives. *Process Saf. Environ. Prot.* **2008**, *86* (5), 328−346.
(7) USEtox model. http://www.usetox.org/ (accessed September 12, 2012).
(8) Rosenbaum, R. K.; Huijbregts, M. A. J.; Henderson, A. D.; Margni, M.; McKone, T. E.; van de Meent, D.; Hauschild, M. Z.; Shaked, S.; Li, D. S.; Gold, L. S.; Jolliet, O. USEtox human exposure and toxicity factors for comparative assessment of toxic emissions in life cycle analysis: sensitivity to key chemical properties. *Int. J. Life Cycle Assess.* **2011**, *16*, 710−727.
(9) Boethling, R. S.; Howard, P. H.; Meylan, W. M. Finding and estimating chemical property data for environmental assessment. *Environ. Toxicol. Chem.* **2004**, *23*, 2290−2308.
(10) U.S. Environmental Protection Agency (2012b). Toxicity Estimation Software Tool (T.E.S.T.). http://www.epa.gov/nrmrl/std/qsar/qsar.html (accessed September 12, 2012).
(11) OSHA PEL. http://www.osha.gov/SLTC/pel/ (accessed September 12, 2012).
(12) Heijungs, R.; Guinée, J.; Huppes, G.; Lankreijer, R. M.; Udo de Haes, H. A.; Wegener, A.; Sleeswijk, A. M. M.; Ansems, P. G.; Eggels, R.; van Duin; de Goede, H. P. *Environmental Life Cycle Assessment of Products. Guide and Backgrounds*; CML (Center of Environmental Science): Leiden, 1992.
(13) Joback, K. K.; Reid, R. Estimation of pure component properties from group contribution. *Chem. Eng. Commun.* **1987**, *57*, 233−247.
(14) Lydersen, A. L. *Estimation of critical properties of organic compounds*, report 3; University of Wisconsin, College of Engineering, Engineering Experimental Station: Madison, WI, 1955.
(15) Constantinou, L.; Gani, R. A new group contribution method for the estimation of properties of pure compounds. *AIChE J.* **2004**, *40* (10), 1697−1710.
(16) Marrero, J.; Gani, R. Group-contribution based estimation of pure component properties. *Fluid Phase Equilib.* **2001**, *183*−*184*, 183−208.
(17) Gani, R.; Harper, P. M.; Hostrup, M. Automatic creation of missing groups through connectivity index for pure-component property prediction. *Ind. Eng. Chem. Res.* **2005**, *44*, 7262−7269.
(18) SimaPro software. http://www.pre-sustainability.com/simapro-lca-software (accessed September 12, 2012).
(19) GaBi software. http://www.gabi-software.com/solutions/life-cycle-assessment/ (accessed September 12, 2012).
(20) Martin, T. M.; Young, D. M. Prediction of the acute toxicity (96-h LC50) of organic compounds to the fathead minnow (*pimephales promelas*) using a group contribution method. *Chem. Res. Toxicol.* **2000**, *14*, 1378−1385.
(21) Casalegno, M.; Benfenati, E.; Sello, G. An automated group contribution method in predicting aquatic toxicity: the diatomic fragment approach. *Chem. Res. Toxicol.* **2005**, *18*, 740−746.
(22) Marrero, J.; Gani, R. A group contribution based estimation of octanol-water partition coefficient and aqueous solubility. *Ind. Eng. Chem. Res.* **2002**, *41*, 6623−6633.
(23) Klopman, G.; Zhu, H. Estimation of the aqueous solubility of organic molecules by the group contribution approach. *J. Chem. Inf. Model.* **2001**, *41*, 439−445.
(24) Kühne, R.; Ebert, R.-U.; Kleint, F.; Scmidt, G.; Schüürmann, G. Group contribution methods to estimate water solubility of organic chemicals. *Chemosphere* **1995**, *30*, 2061−2077.
(25) Martin, T.; Harten, P.; Venkatapathy, R.; Das, S.; Young, D. A Hierarchical Clustering Methodology for the Estimation of Toxicity. *Toxicol. Mech. Methods* **2008**, *18*, 251−266.
(26) U.S. Environmental Protection Agency (2012a). Estimation Program Interface Suite (EPISUITE). http://www.epa.gov/oppt/exposure/pubs/episuite.htm (accessed September 12, 2012).
(27) Isitutio Mario Negri (2012). VEGA. http://www.vega-qsar.eu/contributors.html (accessed September 12, 2012).
(28) Hukkerikar, A. S.; Sarup, B.; Ten Kate, A.; Abildskov, J.; Sin, G.; Gani, R. Group contribution^{+} (GC^{+}) based estimation of properties of pure components: Improved property estimation and uncertainty analysis. *Fluid Phase Equilib.* **2012**, *321*, 25−43.
(29) Seber, G.; Wild, C. *Nonlinear regression*; Wiley: New York, 1989.
(30) Sin, G.; Meyer, A. S.; Gernaey, K. V. Assessing reliability of cellulose hydrolysis models to support biofuel process design − identifiability and uncertainty analysis. *Comput. Chem. Eng.* **2010**, *34*, 1385−1392.
(31) Abramowitz, M.; Stegun, I. A. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*; Dover: New York, 1972.
(32) Mevik, B. H.; Cederkvist, H. R. Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). *J. Chemom.* **2004**, *18*, 422−429.
(33) *Integrated Computer Aided System (ICAS)*, version 15.0; Department of Chemical Engineering, Technical University of Denmark, Lyngby, Denmark, 2012.